Revue québécoise de linguistique

Le modèle *Lstat* : ou comment se constituer une base de données morphologique à partir du Web

Fiammetta Namer

Volume 32, numéro 1, 2003

TALN, Web et corpus

URI : https://id.erudit.org/iderudit/012245ar
DOI : https://doi.org/10.7202/012245ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Namer, F. (2003). Le modèle *Lstat* : ou comment se constituer une base de données morphologique à partir du Web. *Revue québécoise de linguistique*, *32*(1), 85–109. https://doi.org/10.7202/012245ar

Résumé de l'article

L'objectif de cet article est de présenter une méthode d'acquisition, d'organisation et d'interrogation de corpus textuels à partir de données et outils librement récupérables sur le Web (corpus de textes, lexiques, langages de programmation spécialisés, outils d'étiquetage et de lemmatisation). Nous nous intéressons ici plus particulièrement aux préoccupations des chercheurs en morphologie dérivationnelle, en proposant d'intégrer à la méthode présentée un module d'analyse morphologique dérivationnelle qui permet au linguiste de se constituer une base de données lexicale munie d'annotations morphosémantiques (que nous appellerons base de données morphologique, notée BDM). La méthodologie proposée sera illustrée tout au long de l'article par la présentation de **Lstat**, modèle de BDM utilisé avec un lexique de 27,5 millions d'occurrences issu d'archives de presse française en ligne, automatiquement téléchargées entre 2001 et 2002.

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal

LE MODÈLE *LSTAT* : OU COMMENT SE CONSTITUER UNE BASE DE DONNÉES MORPHOLOGIQUE À PARTIR DU WEB

Fiammetta Namer
Université de Nancy II

1. Introduction

On observe une grande variété dans les types d'utilisation du Web à des fins d'analyse linguistique : comme en témoignent les journées consacrées à TALN, Corpus et Web 2002[1], nombre d'auteurs proposent des outils d'extraction automatique de données lexicales spécifiques (gentilés, entités nommées, unités lexicales construites) qui rendent compte de la créativité lexicale dans un domaine donné (Eggert, Maurel et Piton 2002, Fourour et Morin 2002, Hathout et Tanguy 2002); d'autres travaux se servent d'Internet pour l'analyse de contenu (Benamara et Saint-Dizier 2002, Lebarbé 2002, Tazine 2002, Torzec 2002) ou s'interrogent sur la pertinence des requêtes formulées en recherche d'information, et les manières d'y apporter des améliorations (Buvet, Moreau et Silberztein 2002, Emirkanian et Chieze 2002, Fouqueré et Issac 2002); enfin,une expérience de constitution de corpus de référence est en cours de réalisation (Antoniotti et Millon 2002).

L'orientation du travail présenté ici est encore différente : notre objectif est en effet de proposer une méthode d'acquisition, d'organisation et d'interrogation de corpus textuels pour les besoins en linguistique, à partir de données et d'outils librement récupérables sur le Web. La mise en œuvre de cette méthode permet au linguiste, et plus précisément au morphologue, de se constituer une base de données lexicale annotée, à partir des documents en ligne qu'il aura choisis; grâce à un ensemble de requêtes qu'il peut éventuellement compléter, l'utilisateur peut exploiter la base quelles que soient ses compétences informatiques. Illustré tout au long par l'exemple de l'exploitation de corpus journalistiques

1 Le programme et les résumés en sont accessibles à l'adresse : http://www-lli.univ-paris13.fr/.

issus de W3, l'article s'articule autour des sections suivantes : la section 2 motive le dispositif présenté, la section 3 justifie le choix des corpus collectés, la section 4 présente les outils d'acquisition automatique d'information flexionnelle et morphosémantique, la section 5 décrit la base obtenue, les requêtes qui l'exploitent, et analyse un certain nombre de résultats obtenus; enfin, la section 6 dresse le bilan de ce travail, et en explore les perspectives.

2. Web, corpus et morphologie

La possibilité, en morphologie, d'accéder aux documents en ligne et de les exploiter représente un double enjeu. D'une part, pour vérifier ses intuitions, confirmer ses hypothèses, dresser des listes d'exceptions, le chercheur doit confronter les contraintes théoriques des procédés de formation de mots au contenu de corpus dictionnairiques ou textuels. D'autre part, l'étude de ces corpus peut mener à la découverte de constructions lexicales lui suggérant l'existence de nouvelles règles ou conditions d'application de formation d'unités lexicales construites. L'utilisation de corpus est donc centrale pour le travail en morphologie, tant pour l'élaboration que pour la validation d'hypothèses théoriques. En fonction des besoins, on se sert de dictionnaires de langue générale ou spécialisés, et de corpus textuels choisis pour la date de leur création, leur niveau de langue, leur domaine de spécialité, etc.

L'utilisation de corpus de grande taille est donc cruciale, et en cela Internet constitue un réservoir textuel et lexical fabuleux : l'utilisateur morphologue peut y multiplier les études lexicométriques, les découvertes de termes obéissant à des constructions lexicales imprévues; confronter des conditions d'utilisation et de construction lexicales à partir de domaines spécialisés différents; étudier l'impact du registre de langue sur la créativité lexicale… Cependant, on observe que dans le milieu de la recherche en morphologie, le Web reste encore très peu utilisé. En effet, même quand il a identifié avec certitude le type et la localisation de l'arborescence de documents à récupérer, il reste encore au morphologue à résoudre les problèmes de récupération, de traitement, de filtrage, de classement et d'exploitation des données contenues dans ces documents Web.

2.1 Outils existants

Depuis quelque temps, cependant, cette tendance tend à s'inverser : dans les milieux de la recherche francophone en morphologie, notamment, de nombreux outils d'exploitation des corpus en ligne sont mis à la disposition des utilisateurs, créant un pont entre les informaticiens linguistes et la communauté des

morphologues théoriciens. Ainsi, CorpusWeb (Fairon 2000a) télécharge des sites Web et les transforme en corpus, jouant ainsi le rôle d'interface entre le Web et un logiciel de traitement de textes. GlossaNet (Fairon 2000b) est pour sa part un agent de veille qui combine un «aspirateur» Web et un analyseur de corpus, pilotés de telle manière qu'ils soient capables de réitérer automatiquement une série de tâches à intervalles réguliers. WebLex (Heiden et Lafon 2002) est un ensemble d'outils interfacés avec le Web qui permettent d'appliquer des mesures lexicométriques à un texte téléchargé. Parmi les dispositifs qui répondent plus directement aux besoins des morphologues, citons encore Tanguy et Hathout 2002, qui ont conçu et développé Webaffix, outil d'extraction automatique de néologismes construits. Enfin, WaliM (Namer 2003a) est un outil permettant de valider des hypothèses en morphologie par une interrogation automatique de moteurs de recherche au moyen de listes de mots reflétant les hypothèses en question; le résultat de la recherche ramène le cas échéant le contexte phrastique d'apparition des mots testés.

2.2 Lstat : pour une base de données lexicale à la carte

Ce que nous décrivons dans ce qui va suivre n'est pas une application, comme celles mentionnées ci-dessus, mais plutôt une chaîne de traitement réutilisable où se succèdent les méthodes d'acquisition, de filtrage, d'étiquetage des corpus de textes à partir du Web, et de définition du modèle de la base de données relationnelle permettant l'enregistrement des données lexicales ainsi annotées. Ce modèle, baptisé **Lstat**, s'accompagne d'un ensemble de requêtes paramétrables rendant possible l'exploitation, par le morphologue[2], du réservoir de données ainsi constituées.

Comme nous allons le voir, les informations codées sur les entrées lexicales sont d'ordre 1° catégoriel et flexionnel, ce qui permet de relier formes fléchies, lemmes et informations morphosyntaxiques; 2° statistique, ce qui autorise des mesures lexicométriques variées (fréquence, productivité…), en fonction ou non du corpus d'origine, et naturellement d'ordre 3° morphosémantique, de sorte que l'utilisateur peut accéder à la structure des unités lexicales construites, mais également aux contraintes sémantiques imposées par le procédé de construction lexical en jeu.

2 L'exploitabilité de **Lstat** ne se limite pas à la morphologie : les travaux sémantiques en linguistique de corpus peuvent en tirer profit, notamment grâce aux informations sémantiques calculées automatiquement par l'analyseur morphologique DériF (cf. entre autres Namer 2002). C'est cependant dans la perspective d'applications en morphologie dérivationnelle que la version de **Lstat** décrite ici a été constituée.

Dans ce qui suit, nous décrirons tout d'abord (section 3) la méthode mise en œuvre pour l'acquisition des corpus à partir du Web (en l'occurrence des textes issus du *Nouvel Observateur*, de *Science et Avenir*, de *Pour la Science* et de *Challenges*), en détaillant les étapes de traitement préalables à la création de **Lstat**. Ces étapes comportent la traduction en format texte des corpus extraits automatiquement, le filtrage et la correction des erreurs récurrentes, la segmentation, l'étiquetage et la lemmatisation des textes obtenus. Puisque les unités annotées obtenues à l'issue de cette phase sont directement utilisables pour de nombreuses études lexicométriques (calcul de fréquence, mesures de concordance, extractions suivant des patrons syntaxiques …), nous considérons cette phase comme l'étape fondamentale de constitution de la chaîne de traitement.

C'est par l'utilisation de l'analyseur morphosémantique DériF, présenté à la section 4, que le modèle **Lstat** se spécialise en tant qu'outil pour le morphologue : nous verrons comment les corpus annotés s'enrichissent de traits morphologiques fournissant des renseignements tant structurels, catégoriels que sémantiques; puis nous présenterons la façon dont s'organise la BDM permettant à un utilisateur d'accéder à toutes ces informations.

Enfin (section 5), nous montrerons par des exemples de requêtes le cadre d'utilisation de **Lstat** en morphologie. L'utilisateur de **Lstat** peut obtenir des réponses à des questions comme : quels procédés de construction de mot sont les plus représentatifs? quelles unités lexicales morphologiquement complexes ont une base elle-même complexe? quelles sont les unités lexicales morphologiquement complexes attestées dans **Lstat** alors que leur base ne s'y trouve pas? quels sont les hapax[3] (morphologiquement simples et/ou complexes)? Deux résultats de requêtes seront analysés. Perspectives et conclusions constitueront enfin la dernière partie (section 6) de cet article.

3. Acquisition de corpus en ligne et prétraitements

La première phase de la chaîne de traitement se décompose en trois étapes principales. Chacune fait appel à des données ou à des outils de traitement librement disponibles en ligne.

3 Un hapax est une forme qu'on ne rencontre qu'une fois dans la totalité du corpus.

3.1 Choix des corpus, extraction des arborescences, traduction de leur format en texte

La première étape est la collecte de documents textuels en ligne. À partir d'une URL racine, l'ensemble des pages HTML qui en constituent l'arborescence est récupéré automatiquement au moyen de l' «aspirateur» de sites Web standard *wget*, qui a l'avantage d'être utilisable sur toute plate-forme[4]. Ce programme a été intégré dans un script permettant de télécharger automatiquement l'ensemble des pages sélectionnées.

Pour l'expérience relatée ici, notre choix de corpus à collecter à partir du Web s'est porté sur les archives de journaux et revues d'information générale et de vulgarisation. Les raisons en sont : 1° leur grande disponibilité; 2° la quantité importante de documents facilement téléchargeables; 3° le fait que ces documents constituent une masse conséquente de données lexicales homogènes d'un point de vue stylistique; 4° la garantie, en d'autres termes, d'un registre de langue soutenu et cohérent. Ces critères quantitatifs et qualitatifs constituent une première motivation pour le choix des documents. Bien que non représentatifs des ressources documentaires présentes sur Internet, ces contenus d'archives sont assez variés, complets et créatifs du point de vue lexical pour répondre de façon satisfaisante aux besoins d'un morphologue en quête de réponses en corpus. Enfin, et cela constitue la cinquième et dernière raison de notre choix, nous avons délibérément, dans cette expérience, choisi de privilégier un domaine suffisamment vaste en termes de sujets traités (un journal comme *Le Nouvel Observateur*, par exemple, traite de politique, de sciences, de techniques, d'arts, etc.) pour s'apparenter à un genre textuel perçu comme non spécialisé. Comme cela est relaté dans Grabar et Zweigenbaum 2003, l'analyse lexicale et morphologique quantitative de ce type de textes, en opposition, par exemple, à celle faite sur des documents du domaine biomédical, diffère dans les mesures de fréquence, de productivité, et de représentativité des types morphologiques, pour ne citer que ces facteurs.

À partir donc d'archives du site du Nouvel Observateur[5], nous avons collecté l'ensemble des numéros disponibles de l'hebdomadaire *Nouvel Observateur* (au moment de la collecte, 500 numéros étaient disponibles, couvrant la période allant d'août 1993 à mars 2003), de la revue économique *Challenges* (80 numéros, allant d'avril 1997 à février 2003) et du mensuel scientifique

4 cf. http://www.gnu.org/software/wget/wget.html.

5 http://archquo.nouvelobs.com/index.html.

Sciences et Avenir[6] (presque 100 numéros archivés depuis novembre 1996). Au total, la masse de documents colligés correspond, une fois traduite en texte, à un corpus d'environ 27,5 millions d'occurrences.

C'est justement cette traduction en texte seul depuis le format HTML qui constitue la tâche la plus complexe à réaliser, comme le soulignent notamment les auteurs des diverses expériences de téléchargement de corpus de textes à partir du Web relatées lors du colloque TALN, Corpus et Web 2002. Le passage du format HTML en texte brut, qui est ce que Fairon 2000a appelle «transformation d'un site Web en corpus de textes» constitue pour nous une étape nécessaire, mais non suffisante, à la constitution de corpus à partir du Web. L'étape est nécessaire, car nous avons choisi de ne pas mémoriser la structure logique des documents collectés, le but ultime étant l'analyse de données lexicales[7]. Elle n'est pas suffisante, car la récupération du texte brut doit s'accompagner de : 1° la suppression de séparateurs inutiles et de portions de textes, souvent répétitives et relatives à la navigation dans l'arborescence (p. ex. «Retour à la page d'accueil»); 2° la suppression de code Javascript résiduel, de caractères cachés; 3° la détection et le recodage de caractères accentués dans le format d'origine du document; cette tâche a été effectuée en adaptant le programme de recodage de Grabar et Berland 2001; 4° le filtrage de citations en langue étrangère, qui peut être réalisé, comme le proposent Grabar et Zweigenbaum 2003, en reprenant les données de Grefenstette et Nioche 2000, et en en adaptant la méthode.

La difficulté principale de cette première étape de traitement est liée à l'hétérogénéité des formats d'origine des documents en ligne : la nature des balises et le codage des diacritiques, entre autres, varie selon que le HTML est obtenu par traduction au moyen d'un logiciel de traitement de textes ou généré par un éditeur réservé. Il en résulte que les problèmes liés à l'extraction du texte des pages récupérées – et donc les solutions qui doivent y être apportées – diffèrent en général pour chaque arborescence de documents. L'hétérogénéité du format dans lequel sont recueillis les documents en ligne impose également aux filtrages automatiques d'être confirmés presque systématiquement par une vérification manuelle. Actuellement, le contrôle humain complet des résultats de cette première étape serait nécessaire. Il s'agit bien entendu de l'un des inconvénients majeurs à l'utilisation du Web comme source

6 http://www.pourlascience.com de 1997 à 2002; pour chacun des 12 mensuels proposés par an, une sélection de 5 à 8 articles en texte intégral est accessible.

7 On pourra se reporter à Berland 2000 pour la présentation d'une technique alternative de constitution de corpus à partir du Web faisant intervenir une étape de recodage en SGML des balises HTML pertinentes.

de corpus textuels qui nous incite à ne choisir qu'un jeu réduit de types de sources pour une expérience donnée. Une validation automatique par le biais d'une confrontation des occurrences lexicales avec le contenu d'un lexique de grande taille pourrait constituer l'une des pistes possibles pour diminuer l'ampleur de la tâche dévolue à l'opérateur humain.

3.2 Segmentation des unités, étiquetage catégoriel

Le programme TreeTagger (Schmid 1994) entraîné pour le français est en charge de la segmentation et de l'étiquetage grammatical des mots des corpus textuels obtenus. Cet étiqueteur, publiquement disponible à des fins de recherche[8], est indépendant de la langue, et son fonctionnement est probabiliste : il se fonde sur l'utilisation d'arbres de décision et se sert d'un dictionnaire de petite taille. Le système comprend également un module de segmentation; lors de l'étiquetage, le lemme et certaines informations morphoflexionnelles (temps pour les verbes, type du déterminant, etc.) sont calculés. Les résultats de l'étiquetage sont affichés sous forme de triplets réunissant la forme fléchie, la catégorie, le lemme :

Tableau 1
Sortie de Treetagger

Les	DET:ART	le	-c'	ADV	<unknown>
hyperliens	PRO:POS	<unknown>	Heliopolis	NAM	<unknown>

Comme en témoigne le Tableau 1, la valeur arbitraire <unknown> est attribuée aux unités inconnues du catégoriseur («Héliopolis»); la présence de ces mots inconnus («hyperliens») ainsi que les mauvaises segmentations («-c'») peuvent entraîner la production d'erreurs d'étiquetage. Les erreurs les plus fréquentes de TreeTagger concernent les noms propres, rarement reconnus comme tels.

3.3. Calcul de la forme canonique et des traits flexionnels

La dernière tâche de cette phase de prétraitement consiste à lemmatiser les formes catégorisées, après avoir contrôlé, voire rectifié les étiquetages et segmentations erronés. Le programme de lemmatisation doit être capable de

8 cf. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html. Les résultats de TreeTagger sont généralement très bons, et les temps d'exécution très rapides. L'inconvénient principal à son utilisation réside dans le fait qu'il ne tourne que sous Unix/Linux.

plus d'effectuer l'analyse flexionnelle des mots inconnus. Pour réaliser cette tâche, on applique à une forme catégorisée le lemmatiseur Flemm (Namer 2000), qui se sert de règles pour en produire le lemme ainsi que l'ensemble des traits morphoflexionnels calculables hors contexte. Le dictionnaire est réduit à quelques milliers d'exceptions. Puisqu'il est basé sur des règles, ce programme est capable de lemmatiser les mots dits inconnus, et de gérer les analyses multiples. Le résultat produit est de type :

MotFléchi/étiquette:Traits Flexionnels/Lemme(:Famille)

La vérification des résultats de l'étiquetage, qui précède la lemmatisation proprement dite, se déroule en deux temps : 1° décollement des ponctuations en début et en fin de mot, et s'il y a reconnaissance de la séquence graphique obtenue, affectation de la catégorie grammaticale correspondante; 2° évaluation de la validité de la catégorie affectée par l'étiqueteur en fonction de la terminaison du mot en présence : en cas d'incompatibilité, la catégorie choisie par Flemm est à la fois la plus vraisemblable du point de vue de la graphie du mot, et la plus fonctionnellement proche de la catégorie rejetée. En reprenant l'échantillon du Tableau 1, la vérification affecte les entrées «–c'» et «hyperliens» : 1° une fois la séquence «c'» isolée du tiret, elle est reconnue comme pronom (PRO); 2° la classe fermée des pronoms possessifs (PRO : POS) n'incluant pas «hyperliens», ce dernier est recatégorisé nom commun (NOM). Le résultat de l'application de Flemm sur l'échantillon du Tableau 1 est donné dans le Tableau 2, qui montre la lemmatisation de mots inconnus («Héliopolis») éventuellement après réétiquetage («hyperliens», «c'»), et ajout de traits flexionnels (le nombre (p)luriel sur les entrées de «les» et «hyperliens») :

Tableau 2
Sortie de Flemm

Les	DET:(ART):_:p: le	-c'	PRO	ce
hyperliens	NOM:_:p hyperliens	Heliopolis	NAM	Heliopolis

4. Analyse morphosémantique des unités lexicales : annotations pour le morphologue

À la sortie de Flemm, les annotations produites sont suffisantes pour alimenter une base de données lexicales exploitable en linguistique de corpus. L'ajout d'un analyseur morphosémantique comme maillon supplémentaire

de la chaîne constituée jusqu'ici va permettre à l'utilisateur de formuler de nouvelles requêtes concernant par exemple les familles morphologiques des unités lexicales construites, la productivité d'un affixe en fonction du corpus d'origine, les dérivés apparaissant avec leur base, la complexité morphologique moyenne des mots utilisés dans tel ou tel type de document, la nature sémantique des liens que tissent une base et son dérivé, ainsi que d'autres traits de sémantique lexicale.

L'objet de cette section est la description d'un tel analyseur. Il s'agit du système DériF (Dérivation en Français), développé dans le cadre du projet MorTAL (Hathout, Namer et Dal 2002). DériF s'oppose dans sa conception ainsi que dans ses objectifs à une vision antérieure du traitement automatique de la morphologie fondée sur un modèle à deux niveaux, suivant une vision concaténatoire de la formation des unités lexicales complexes. Dans cette perspective théorique, dont la première implémentation, due à Karttunen 1983, est décrite en détail dans Sproat 1992[9], la construction du lexique est perçue comme fondamentalement morphématique, dissociative (le mécanisme de désaffixation est distinct de l'ensemble des procédés d'affectation du sens) et linéaire. Comme le soulignent entre autres Gruaz, Jacquemin et Tzoukermann 1996, un modèle à deux niveaux pour la dérivation du français est surgénératif (il accepte des constructions linguistiquement invalides), tout en ne permettant d'exprimer ni la hiérarchie des procédés impliqués, notamment lors de l'alternance préfixe/suffixe, ni par conséquent les relations sémantiques induites par chaque procédé.

Nous nous situons, à l'inverse, dans la mouvance des hypothèses théoriques émises au départ par Corbin 1987, qui défend, elle, une morphologie lexématique[10] et associative, où tout type de procédé morphologique (suffixation, préfixation, conversion, composition) met en jeu des règles qui imposent à la base sélectionnée et au dérivé construit des contraintes à la fois sémantiques, phonologiques, catégorielles, etc. Un système modélisant ces résultats théoriques se doit de prendre en compte les propriétés fondamentalement sémantiques des règles morphologiques, qui s'organisent selon une hiérarchie linguistiquement motivée, subordonnée à des exceptions que le système doit être à même de prendre en compte. C'est ce que réalise DériF en effectuant l'analyse morphologique complète du lemme étiqueté qui lui est fourni en entrée, jusqu'à

9 Se reporter à Fradin 1994 pour une présentation générale de l'approche à deux niveaux, et à Daille, Fabre et Sébillot 2002 pour un panorama des applications de la morphologie computationnelle.

10 Voir, entre autres, Fradin 2003 pour une comparaison des points de vue morphématique contre lexématique de la morphologie dérivationnelle.

l'obtention d'une unité lexicale morphologiquement indécomposable. Les principes linguistiques qui servent de support à l'algorithme d'analyse et qui s'inspirent de la théorie élaborée à l'origine par D. Corbin modélisent et implémentent également les résultats présentés entre autres dans Aliquot-Suengas 1996; Amiot 1997; Corbin 1991, 1997, 2000a, 2000b; Dal 1997a, Fradin 2003; Plénat et Roché 2000; Temple, 1996).

4.1 Résultats produits

À ce jour, DériF effectue l'analyse morphologique et sémantique complète des unités lexicales suffixées par *-able, -ité, -et(te), -is(er), -ifi(er), -eur, -ment, -tion, -oir*, préfixées par *dé-, in-, re-, a- en-*, ou obtenues par conversion adjectif → verbe[11]. Comme cela est souligné dans Namer 2002, 2004a et illustré dans la série d'exemples de la Fig. 1, le système calcule l'arbre d'analyse d'un lemme étiqueté, cet arbre étant repris sous forme de famille ordonnant l'ensemble des bases successives reconnues par l'analyseur; la troisième partie du résultat consiste en la représentation en langage naturel de la relation sémantique que tisse l'«input» avec sa base; la quatrième et dernière partie est l'ensemble des traits sémantiques automatiquement acquis et affectant la base et/ou le dérivé en fonction des contraintes exercées par le procédé de construction (ex. 1). Les caractéristiques principales de DériF sont 1° la **récursivité** : l'analyse d'un mot est réitérée jusqu'à l'obtention d'une base non analysable (ex. 2); 2° le **traitement hiérarchisé** de diverses opérations de construction (suffixation, préfixation, conversion, composition), en fonction de la portée respective des procédés en présence. Ainsi, les exemples (3a) et (3b) illustrent l'ordre inverse d'analyse de la préfixation et la suffixation; 3° la **gestion des analyses ambiguës** : DériF manipule des listes de données et de résultats, et est donc capable de générer autant de solutions qu'il y a d'ambiguïtés éventuelles dans l'analyse d'un mot construit (ex. 4); 4° le **traitement des néologismes** : de par sa conception, DériF est capable d'analyser des mots inconnus, possibles et non attestés (ex. 5).

11 Le procédé de conversion consiste à construire une unité lexicale d'une catégorie C1 à partir d'une unité lexicale base de catégorie C2, sans l'entremise de matériel lexical supplémentaire. Les conditions portant sur C1, C2, et sur les caractéristiques sémantiques de la base et du mot construit contraignent à la fois les types de conversion possibles (ainsi, verbe → adjectif est impossible), et l'orientation des conversions autorisées verbe → nom implique que le nom est abstrait, p. ex. *voler → vol*, alors que dans nom → verbe, le nom est souvent l'instrument du procès décrit, p. ex.. *balai → balay(er)* (cf. entre autres Kerleroux 1996, 1997, 2000 pour une description approfondie de la conversion).

1) **continuité, NOM** ==> [[continu ADJ] ité NOM] (continuité/NOM, continu/ADJ) :: «faculté d'être continu»

continuité (abstrait,propriéte,xxx)

continu (xxx,xxx,xxx,prédicatif)

2) **explicabilité, NOM** ==> [[[expliquer VERBE] able ADJ] ité NOM] (explicabilité/NOM, explicable/ADJ, expliquer/VERBE) :: «faculté d'être explicable»

explicabilité (abstrait,propriété,xxx)

explicable (xxx,inhérent,exogène,prédicatif)

expliquer (xxx,transitif, [agent,thème])

3a) **introuvable, ADJ** ==> [in [[trouver VERBE] able ADJ] ADJ] (introuvable/ADJ, trouvable/ADJ, trouver/VERBE) :: «non trouvable»

3b) **désossable, ADJ** ==> [[dé [os NOM] VERBE] able ADJ] (désossable/ADJ, désosser/VERBE, os/NOM) :: «qu'on peut désosser»

désossable (xxx,inhérent,exogène,prédicatif)

désosser (xxx,transitif, [agent,thème])

4) **desservir, VERBE** ==> [dé1 [servir VERBE] VERBE] (desservir/VERBE, servir/VERBE) :: «(Enlever ce qui a pour effet de | Annuler l'état lié au procès) de servir»

desservir, VERBE ==> [dé2 [servir VERBE] VERBE] (desservir/VERBE, servir/VERBE) :: «Cesser de servir; servir fortement, intensément, jusqu'au bout, au loin»

5) **benladenisation, NOM** ==> [[[Benladen NPR] is(er) VERBE] tion NOM] (benladenisation/NOM, benladeniser/VERBE, Benladen/NPR) :: «action ou résultat de benladeniser»

benladenisation (abstrait,action/résultat,xxx)

benladeniser (causatif, transitif, [cause,thème])

Fig. 1 : Exemples d'analyses de DériF

4.2 Fonctionnement de DériF

Le fonctionnement de DériF est illustré dans la Fig. 2, par l'analyse de l'adjectif *incontournable*. Le constituant fondamental de l'analyseur est un ensemble de deux moteurs qui choisissent les procédures d'analyse à déclencher en fonction de la graphie et de la catégorie du lemme qu'ils examinent. Le principe d'activation de chaque moteur est le suivant : le lemme catégorisé constitue l'entrée du programme. Aucune autre information n'est disponible ni requise. Le premier moteur s'active. Il a pour tâche de déceler dans la terminaison de l'input une forme suffixoïdale compatible avec l'étiquette grammaticale de celui-ci : dans la Fig. 2, la séquence identifiée est *–able*, catégoriellement compatible avec ADJ. Le moteur appelle alors la fonction Fsuf d'analyse spécifique pour le type morphologique pressenti, à savoir ici celui des adjectifs suffixés par *–able*.

Fsuf vérifie tout d'abord que son input n'est pas le résultat de l'application d'un préfixe sur la base suffixée par *–suf*. Cette étape de vérification n'examine que les préfixes catégoriellement compatibles avec les lemmes en *-suf* : Ainsi, pour Fable, la liste des préfixes pouvant porter sur la base suffixée contient *in-, auto-, super-, hyper-*… mais exclut *re-, dé-, pré-* …[12] C'est également à ce stade que va se déclencher la recherche de lemmes morphologiquement ambigus (p.ex. *inversable, imposable, importable,* qui peuvent donner lieu à deux analyses possibles). L'identification du préfixe *in-* appliquée à la base *contournable* est confirmée par la fonction Fpref1. La fonction calcule également la relation sémantique que tisse l'input avec *contournable* (à savoir «non contournable»); elle renvoie ensuite son résultat à Fable. La décomposition base/suffixe consiste ensuite à reconstituer le mot-base du lemme suffixé, en appliquant si besoin est les règles d'allomorphie à la séquence obtenue par désuffixation (notée Z dans la Fig. 2, et ayant ici pour valeur : *contourn*). Le résultat Z' (ici le verbe *contourner*) est renvoyé finalement au moteur. À Z' s'ajoutent les listes de traits syntaxicosémantiques que le suffixe *–able* affecte à sa base verbale (indiquant qu'elle est transitive) et à l'adjectif construit (indiquant qu'il exprime une propriété latente et non acquérable).

Tant qu'il reconnaît un suffixe dans la terminaison du résultat qu'il réceptionne, le premier moteur réitère l'ensemble des tâches décrites ci-dessus. Dans le cas du radical *contourn-* du verbe *contourner,* par contre, aucune forme suffixoïdale n'est détectée, et la récursion s'achève : le second moteur est ap-

12 Cette liste de préfixes n'est pas exclusive, et pour chacun d'eux le programme répertorie un certain nombre d'exceptions : ainsi, *superposable* n'est pas construit sur *posable*, et à l'inverse, *défavorable* a pour base *favorable*.

pelé. Sa tâche est de rechercher si le lemme, reconnu comme non suffixé, est le résultat d'opérations, éventuellement répétées, de préfixation ou de conversion. Structurellement, le verbe *contourner* semble être analysable comme construit par préfixation par *co-* sur la base verbale *tourner*. Cependant, le sens construit d'un verbe déverbal formé par *co-* (étymologiquement issu du latin «cum» = *avec*) exprime l'application conjointe ou parallèle d'un procès sur deux objets (*codiriger, coadapter, coagir*…). *Contourner* ne peut être glosé au moyen d'une telle relation. Dans ce décalage entre complexité apparente de la structure et non-compositionnalité du sens, la primauté de la sémantique sur le formel et le catégoriel (cf. Corbin 2000c; Dal 1997b) nous amène à considérer *contourner* comme un verbe non construit[13]. Le second moteur, alerté par la liste d'exceptions pertinente, n'active donc aucune fonction d'analyse. Le résultat est alors affiché sous la forme décrite en 4.1.

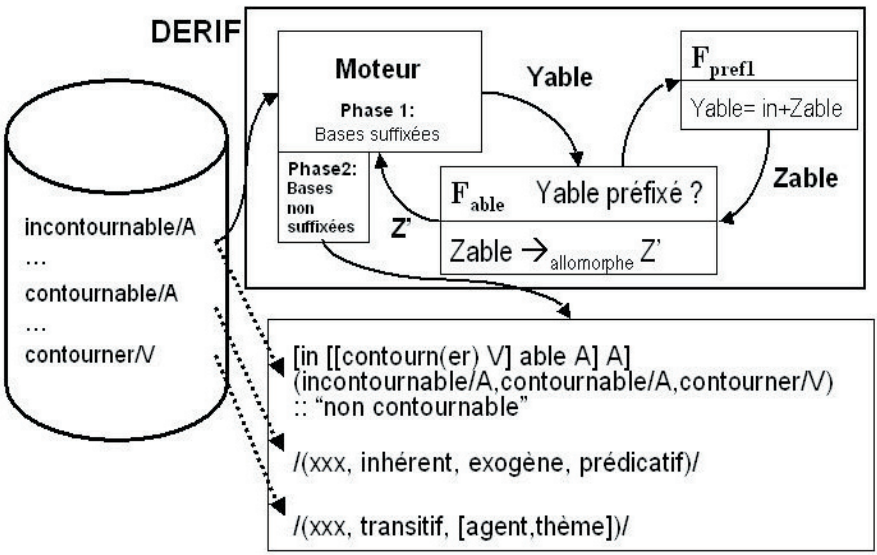

Fig. 2 : Analyse par DériF de l'adjectif *incontournable*

Parce qu'il est en cours de développement, DériF est le seul outil présenté dans cet article auquel il est encore impossible d'accéder sur Internet. Cette situation devrait changer d'ici fin 2004, date à laquelle DériF sera publiquement disponible à des fins de recherche.

13 Cette analyse trouve confirmation dans l'examen de l'étymologie de ce verbe, relevée dans Rey 1998, et qui indique que c'est l'étymon latin du verbe qui est construit morphologiquement : «(…) est issu du latin populaire *contornare*, (…) de *cum* (→ co) et *tornare* (→ tourner)».

Parmi les procédés de construction lexicale non encore pris en compte par DériF, citons les suffixes formateurs d'adjectifs dénominaux. Pour pallier partiellement cette lacune temporaire, nous avons associé à DériF un ensemble de règles obtenues par apprentissage à partir de corpus spécialisés de grande taille. La technique d'acquisition de ces règles est décrite dans Grabar et Zweigenbaum 2003. Contrairement à ce que fait DériF, cet ensemble additionnel de règles ne réalise qu'une décomposition formelle des adjectifs construits par suffixation sur des bases nominales. Les suffixes utilisés sont principalement *–al* (et sa variante allomorphique *–el*), *-ais*, *-aire, -ique, -é, -eux, -ien, -ois*.

5. Mise en œuvre de Lstat et interrogation

Une fois réalisées sur les occurrences des corpus les annotations catégorielles, morphosyntaxiques, flexionnelles et morphosémantiques, ces données sont organisées sous forme de base de données (§5.1). Après description du modèle (§5.2) et présentation des requêtes types que l'utilisateur peut effectuer (§5.3), l'usage de **Lstat** est illustré de quelques résultats (§5.4).

5.1. Formatage des corpus annotés

À l'issue de la phase d'analyse morphosémantique réalisée par DériF, les corpus annotés subissent un reformatage automatique afin de faciliter l'alimentation par ces corpus de la BDM **Lstat**. Cette réécriture est illustrée, dans la Fig. 3, avec l'exemple de l'entrée de *incontournable* issue du document *Pour la Science.* Pour un même document source, toutes les occurrences d'un même triplet forme fléchie/catégorie/lemme sont regroupées de manière à mémoriser le nombre de ces occurrences, et la position de chaque occurrence dans le document d'origine. Cette indication est précieuse, parce qu'elle sert à resituer le mot dans son contexte, et parce qu'elle permet d'attribuer à la BDM les fonctions de concordancier. La Fig. 3 montre que la forme masculin singulier du lemme adjectival *incontournable* est un hapax dans *Pour la Science*, son unique réalisation constituant la 14 132^{e} occurrence du texte. L'analyse par DériF fournit la complexité morphologique du lemme du triplet, calculée à partir du nombre d'éléments dans la famille d'analyse de ce lemme (cf. section 4.1) : comme l'indique la Fig. 2 ci-dessus, la famille de *incontournable* est la liste *(incontournable/ADJ, contournable/ADJ, contourner/VBE),* sa complexité morphologique vaut donc 3. À partir de la représentation hiérarchisée (crochetée) de l'analyse sont reconstituées ensuite les différentes étapes, séparées par

une virgule, reliant le lemme à l'unité lexicale indécomposable obtenue par DériF. Chaque étape mémorise 1° la règle de formation en jeu, c'est-à-dire dans l'ordre (Fig. 3) la catégorie de la base, le nom et le type du procédé morphologique mis en jeu, la catégorie du dérivé, et 2° le lemme-base et sa catégorie. Par exemple, la première étape de l'analyse de *incontournable* met en jeu la règle ADJ_{base}/in/pre/$ADJ_{der.}$, qui aboutit à l'adjectif base *contournable*.

```
14132  1     incontournable   ADJ  incontournable
3,ADJ/in/pre/ADJ+contournable/ADJ,
VBE/able/suf/ADJ+contourner/VBE
```

Fig. 3 : Représentation de l'occurrence de *incontournable* dans *Pour la Science*

Dans la version actuelle du modèle ne sont intégrés ni les traits flexionnels calculés par Flemm, ni les informations sémantiques (relation base/dérivé, et liste de traits syntaxicosémantiques sur base et/ou dérivé) obtenues par DériF; c'est pourquoi ces informations ne sont pas reproduites dans la reformulation illustrée dans la Fig. 3. Par contre, pour une entrée correspondant à un lemme morphologiquement ambigu, on représente l'ensemble des interprétations de ce lemme que DériF a identifiées, séparées par «|». L'entrée de déposera, dans la Fig. 4, en est un exemple : le triplet *déposera/VBE/déposer* apparaît trois fois dans *Pour la Science*, aux positions indiquées; *déposer* est morphologiquement ambigu : soit il s'analyse comme l'inverse de *poser* (au moyen du préfixe *dé1-* «privatif»), ce qu'on rencontre p. ex. dans «*les ouvriers ont déposé le papier peint*»; soit il s'analyse comme un perfectif (au moyen du préfixe homographe *dé2-*), comme l'atteste l'exemple «*Max a déposé les copies dans le casier*». Enfin, quelle que soit l'analyse, la complexité morphologique de *déposer* vaut 2 :

```
10232:10307:13351 3     déposera    VBE    déposer
2,VBE/dé1/pre/VBE+poser/VBE| 2,VBE/dé2/pre/VBE+poser/VBE
```

Fig. 4 : Représentation des occurrences de *déposera* dans *Pour la Science*

5.2 Modèle **Lstat**

L'organisation des informations calculées par la chaîne de traitement est compatible avec le modèle de base de données relationnel **Lstat**, où interagissent les tables illustrées dans la Fig. 5. Les tables principales correspondent aux

trois niveaux d'interrogation de la base : lexical et contextuel (table Usages), morphoflexionnel (table Flexions) et morphodérivationnel (table Analyses) :

1° Usages pour un corpus (Corpora) donné, on identifie au moyen d'un Compteur un certain nombre d'occurrences. Chaque occurrence est reliée à une Localisation, c'est-à-dire une des positions du triplet dans le corpus.

2° Flexions : identifie chaque occurrence par le triplet Forme Fléchie, Catégorie, Lemme.

3° Analyses : associe tout lemme morphologiquement construit du triplet ci-dessus à une ou plusieurs interprétations. Chacune d'entre elles, (une Analyse), est un enchaînement d'Étapes reliant le lemme à l'unité lexicale indécomposable calculée par DériF, via une succession de bases intermédiaires : chaque Étape réunit une base (Bases) et un procédé de construction de mot (Procédés Morphologiques); le procédé se décompose, lui, en l'opération morphologique elle-même (*-able, re-* …) (Opérations Morphologiques) et son type (suffixe, préfixe, mais aussi conversion ou composition) (Types Opérations Morphologiques).

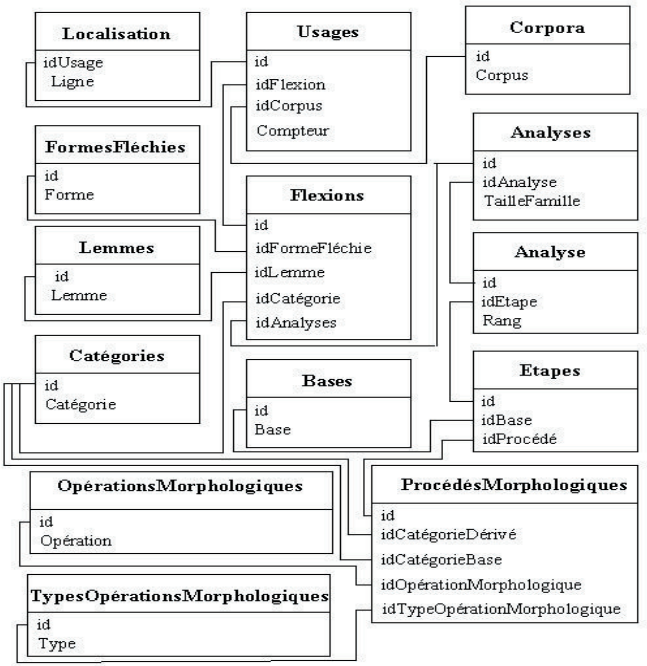

Fig. 5 : Modèle **Lstat**

5.3 Exemples de requêtes

Les documents collectés sur le Web (section 3.1), annotés morphologiquement à la suite de la chaîne de traitement (sections 3.2, 3.3 et 4) puis reformulés comme indiqué dans la section 5.1, servent d'entrée à la création de la BDM générée selon le modèle **Lstat** (section 5.2). Comme nous le verrons en conclusion, d'autres documents sources, toujours issus d'Internet, ont par ailleurs été utilisés pour la réalisation d'autres BDM sur ce même modèle.

Un système composé d'environ 80 requêtes prédéfinies accompagne la création de la BDM. Ces requêtes sont soit formulées en MySQL, soit conçues au moyen de scripts Perl paramétrés, laissant à l'utilisateur le soin d'y apporter les précisions souhaitées. Cet ensemble de scripts traduit une partie des interrogations qu'un utilisateur peut vouloir formuler en morphologie; nous en proposons ci-dessous une liste non exhaustive.

1° Quelle est la fréquence d'une forme fléchie? d'un lemme? de l'ensemble des formes fléchies et/ou des lemmes? Quelle est la fréquence moyenne d'un procédé de construction morphologique, soit quelle est la proportion de lemmes (formes fléchies de lemmes) obtenus par le biais de ce procédé par rapport à l'ensemble des lemmes (formes fléchies) de la même catégorie grammaticale?

2° Quelle est la liste des formes fléchies (lemmes) obtenues au moyen d'un procédé de construction? Quelle en est la liste, pour l'ensemble des procédés de construction répertoriés dans la BDM?

3° Quels sont les lemmes construits dont la base est présente dans le même corpus d'origine? Parmi ces lemmes construits, quels sont les hapax? À l'inverse, quelles bases de lemmes construits ne sont pas répertoriées dans la BDM?

4° Quelle est la productivité quantitative d'un procédé de formation P1, c'est-à-dire quelle est la propension de P1 à créer des néologismes? (cf. Baayen 2001; Baayen et Lieber 1991) Quels sont les procédés de construction de mots les plus productifs dans la langue générale? dans une langue de spécialité donnée, et identifiée par un ou des corpus sources sélectionnés?

5° Quelle est la proportion de lemmes morphologiquement construits, en fonction de leur catégorie grammaticale? en fonction de leur corpus d'origine? Y a-t-il des lemmes de complexité morphologique supérieure à trois?

6° Quels sont les noms qu'on rencontre dans le contexte phrastique d'un verbe V, en fonction des corpus d'origine?

7° Pour tout adjectif dénominal A, quels sont les noms qui régissent A?

8° Quels sont les adjectifs dénominaux construits par des suffixes concurrents à partir d'une même base?

9° Quel est le contexte d'apparition dans le corpus source des lemmes construits morphologiquement ambigus?

5.4 Exemples de résultats

À titre d'illustration, nous présentons le résultat de deux requêtes exécutées successivement sur les corpus du *Monde* et de *Challenges*.

Tout d'abord, la Fig. 6 contient la réponse partielle (faute de place, seul le début de la réponse est affiché) à la requête «Quels sont les hapax construits dont la base est elle-même complexe?» : cette réponse témoigne du fait que les néologismes construits sur une base morphologiquement construite sont rares (0,01 % des occurrences à partir des archives du *Monde*); ensuite, elles fournissent une indication sur l'interaction entre création lexicale et morphologie dans le genre textuel illustré par ce corpus : dans l'échantillon, la plupart des néologismes, qui sont des noms, correspondent à des (multi-)compositions spontanées (*acteur-animateur-chanteur*), souvent générées en rafales. Le résultat d'une requête complémentaire, demandant la localisation des noms commençant par *acteur-,* a confirmé la proximité des quatre premiers termes listés dans la figure ci-dessous. Parmi les autres néologismes observés, on remarque des noms convertis à partir d'une base adjectivale préfixée par *anti-* et très souvent elle-même complexe. On peut peut-être expliquer cette propension à la formation de néologismes en *anti-* par l'aspect «investigatif» des réflexions proposées dans un journal d'information comme *Le Monde*.

```
acteur-animateur-chanteur, NOM; acteur-réalisateur,
NOM; acteur-danseur, NOM; acteur-danseur-bruiteur-
éructeur, NOM; acteur-danseur-chanteur, NOM;
acteur-diffuseur, NOM; actualisable, ADJ; agent-
manager-organisateur, NOM; aide-conducteur, NOM;
aide-déménageur, NOM; allocataire-moniteur, NOM;
alphabétiseur, NOM; alternateur-démarreur, NOM;
amortalité, NOM; animateur-producteur, NOM; animal-
fleur, NOM; anti-douceur, NOM; anti-dépresseur,
NOM; anti-fumeur, NOM; anti-invalidité, NOM; anti-
mendicité, NOM; anti-peur, NOM; anti-tumeur, NOM;
anticasseur, NOM; antihypertenseur, NOM
```

Fig. 6 : Hapax de complexité morphologique > 2

La Fig. 7 fournit, elle, un extrait de la réponse à la question Quelle est la liste des séquences de lemmes de type 'Nom Adj' ou Adj est dénominal? Cette requête vise à rechercher les variantes de termes où le nom tête est modifié par un adjectif relationnel, et à typer morphologiquement ces adjectifs. Les couples collectés lors de cette requêtes ont une double utilité : 1° ils servent à identifier en contexte les adjectifs morphologiquement construits à interprétation relationnelle, et les suffixes les plus fréquemment mobilisés pour leur formation; 2° ils constituent souvent un indice pour l'aide à l'indexation du corpus (en tant que variantes de termes) : on remarque en effet que dans de nombreux cas, la séquence NA est caractéristique du vocabulaire économique propre à *Challenges*. Une requête complémentaire, d'ailleurs, confirme cette impression : les couples NOM ADJsuf terminologiquement compatibles avec le domaine économique sont ceux dont le nombre d'occurrences est le plus élevé. Dans la Fig. 7, pour des raisons d'espace, seuls les termes du vocabulaire de la finance sont affichés. Ils constituent plus d'un tiers de l'intégralité de la réponse.

```
argent/NOM américain/ADJ, action/NOM agro-
alimentaire/ADJ, valeur/NOM américain/ADJ, fusion/
NOM bancaire/ADJ, record/NOM boursier/ADJ, air/NOM
boursier/ADJ, abondance/NOM boursier/ADJ, compagnie/
NOM bancaire/ADJ, inflation/NOM brutal/ADJ, produit/
NOM capital/ADJ, salaire/NOM créatif/ADJ, argent/NOM
européen/ADJ, chronique/NOM économiste/ADJ, homme/NOM
économiste/ADJ, placement/NOM européen/ADJ, société/
NOM européen/ADJ, volcan/NOM financier/ADJ, palmarès/
NOM financier/ADJ, harmonisation/NOM fiscal/ADJ,
compagnie/NOM financier/ADJ, action/NOM focal/ADJ,
valeur/NOM fondamental/ADJ, patron/NOM fortuné/ADJ
```

Fig. 7 : Échantillon des séquences NOM ADJsuf dans *Challenges*

6. Perspectives et conclusion

Le modèle **Lstat** a été utilisé à ce jour pour exploiter deux ensembles de documents à partir du Web : les corpus journalistiques de presse quotidienne et hebdomadaire, qui ont illustré cette présentation, ont servi a évaluer la validité en français des résultats expérimentaux obtenus dans Krott, Baayen et Schreuder 1999, à propos de l'interaction entre fréquence, productivité et représentativité

de procédés morphologiques s'appliquant à des mots construits. Les conclusions sont rapportées dans Namer 2003b. À ce jour, **Lstat** est alimenté avec des corpus contenant des documents du domaine biomédical, dans le cadre de projets visant à intégrer les données francophones du domaine, lexicales et morphologiques à la base de connaissances UMLS de la «National Library of Medicine» (NLM)[14].

6.1 **Lstat** et les bases de données en biomédecine

Actuellement, **Lstat** est utilisé pour mettre en commun les ressources et outils nécessaires au consortium UMLS[15] dans la constitution d'un lexique francophone de la langue médicale : les sources (thésaurus, lexiques, bases de connaissances, textes scientifiques/pédagogiques, comptes rendus hospitaliers, etc.) sont en effet essentiellement collectées sur le Web (pour une description du projet, cf. Zweigenbaum et coll. 2003a, 2003b). Dans le but de disposer à terme de toutes les informations requises, il est prévu une extension de **Lstat** pour que les champs suivants, pertinents en terminologie médicale, soient pris en compte : la langue (latin, grec, français, français canadien, français suisse…), le type d'entrée (mot simple, mot construit, affixe…), les informations morpho-flexionnelles (ces traits sont déjà disponibles à la sortie de Flemm, cf. section 3.3), et les renseignements d'ordre extralinguistique tels que l'auteur et la date d'un nouvel ajout dans la BDM.

La dernière évolution de **Lstat** concerne les informations d'ordre syntaxi-cosémantique que DériF est capable d'acquérir automatiquement. L'intégration de ces traits et relations sémantiques se fait à l'occasion du démarrage du projet VUMeF[16]. Parmi les objectifs de VUMeF (cf. Darmoni et coll. 2003), il est en effet prévu d'élargir le contenu du lexique UMLF : les termes multilexicaux y seront intégrés, ainsi que les informations nécessaires à l'identification de synonymes, hyperonymes, antonymes, etc. En ce sens, les traits et relations acquises automatiquement par DériF peuvent jouer un rôle non négligeable :

---

14 Le projet «Unified Medical Language System» (UMLS) de la NLM développe et distribue des données et des programmes libres de droits pour la gestion de lexiques biomédicaux. Ces ressources sont massivement utilisées en recherche sur Internet, bibliographie, TALN, et aide à la décision par les spécialistes du domaine. À ce jour, ces connaissances sont développées essentiellement pour l'anglais.

15 Projet ACI n° 02C0163, 2002-2004, piloté par P. Zweigenbaum, STIM/DSI, AP, Hôpitaux de Paris.

16 Ce projet, qui vient d'être adopté par le Ministère de la Recherche dans le cadre du programme RNTS (Réseau National des Technologies de la Santé), est piloté par la société Vidal, et coordonné par S. Darmoni, L@STICS, CHU de Rouen

par exemple, c'est par le biais de leur analyse morphologique, produisant dans les deux cas la glose «relatif à estomac» qu'on détecte le lien de synonymie existant entre *stomacal* et *gastrique*. Le même mécanisme permettra de conclure à l'équivalence référentielle de *anthropoforme* et *anthropomorphe*; ou encore de classer *phlébite* («inflammation d'une veine») comme un sous-type d'*angéite* («inflammation d'un vaisseau»). D'autres mécanismes, décrits dans Namer 2004b, et Namer et Zweigenbaum 2004, permettent enfin de tisser automatiquement des liens de cohyponymie, par exemple entre *gastralgie* et *hépatalgie*.

6.2 Conclusion : de l'utilisation du Web en linguistique?

Nous avons vu dans cet article comment une chaîne de traitement pouvait être constituée, essentiellement à partir d'outils disponibles en ligne, pour exploiter des documents textuels collectés sur le Web à des fins de recherche orientée en morphologie. La méthodologie sous-jacente à cette chaîne de traitement, aboutissant au modèle de BDM **Lstat**, a été mise en oeuvre dans le cadre de plusieurs projets sommairement décrits à la section 6.1.

Lstat n'est bien sûr qu'un prototype, perfectible à bien des égards : tout d'abord, une interface est souhaitable, pour faciliter à l'utilisateur morphologue non-informaticien la conception de nouvelles requêtes. Elle est actuellement en cours de réalisation. Ensuite, les outils d'acquisition et d'annotation lexicales sont parfois incomplets : c'est le cas de DériF, qui ne prend pas encore en compte les mots composés savants. Ce travail est en cours dans le projet UMLF; on a vu aussi que les outils d'annotation peuvent engendrer des erreurs : TreeTagger gère très mal les entités nommées, ainsi que les mots étrangers. Pour limiter la répercussion sur la lemmatisation et la dérivation des erreurs d'étiquetage que cette non-reconnaissance pourrait engendrer, il est envisagé dans l'UMLF une phase de précodage qui étiquettera les mots latins en fonction d'une liste résultant d'apprentissages antérieurs.

Enfin, la difficulté principale dans l'utilisation du Web comme réservoir de ressources textuelles réside dans le téléchargement et le nettoyage des pages : la validité des réponses fournies par la BDM aux requêtes linguistiques est avant tout subordonnée à la qualité des documents HTML collectés. Or, ceux-ci présentent souvent des formats incohérents, des codages de caractères obéissant à des normes variées, des traces de césures automatiques, des erreurs de typographie… La diversité des sources est dans ce domaine un facteur supplémentaire de risque d'erreurs, et seule une vérification humaine, longue, ennuyeuse mais indispensable, peut garantir la qualité des textes extraits des pages issues du Web.

Cependant, malgré cet inconvénient, il est indubitable que les données du Web constituent une ressource exploitable en linguistique : nous pensons en effet avoir apporté dans cet article des arguments corroborant cette hypothèse : 1° des masses de documents en perpétuelle évolution y sont à notre portée dans toutes les langues, couvrant tous les domaines, relevant de tous les registres; 2° des outils libres de droits, et le plus souvent en ligne, sont à notre disposition pour constituer une chaîne de traitement pour la collecte et l'annotation de ces documents; 3° grâce à ces outils, le matériel lexical contenu dans ces documents a pu être organisé sous la forme d'une base de données lexicales dont le modèle est extensible; 4° le champ d'investigation des morphologues est alors potentiellement infini, et le nombre de requêtes «prêtes à l'emploi» leur permet déjà d'en tirer un nombre non négligeable d'informations.

Par l'exposé de ces points, nous espérons avoir apporté une réponse positive à la question que se pose ce numéro thématique, à propos de l'exploitabilité des données et outils du Web dans le cadre d'applications en linguistique[17].

Références

Aliquot-Suengas S. 1996 *Référence collective / sens collectif. La catégorie du collectif dans les noms suffixés du lexique français*, Thèse de doctorat, Université de Lille III.

Amiot, D. 1997 *L'antériorité temporelle dans la préfixation en français*, Villeneuve d'Ascq, Presses Universitaires du Septentrion.

Antoniotti, M. et Ch. Millon 2002 «Une expérience de constitution d'un corpus de référence du français contemporain à partir du Web», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse.

Baayen, H. 2001 *Word Frequency Distributions*, Dordrecht, Kluwer.

Baayen, H. et R. Lieber 1991 «Productivity and English derivation : a corpus-based study», *Linguistics* 29-5: 801-843.

Benamara F. et P. Saint-Dizier 2002 «Analyse et exploitation des données du Web par un extracteur dynamique de connaissances», communication présentée au colloque TALN, Corpus et Web 2002*,* Villetaneuse. [texte ici même]

Berland, S. 2000 *Constitution de corpus à partir du Web pour l'acquisition terminologique : une expérience*, mémoire de DESS Ingénierie Multilingue, Paris, INALCO.

17 **Lstat** est disponible sur demande, mais ne comprend actuellement que les outils présentés dans les sections 3 et 5. Pour l'instant, DériF n'est pas accessible : dès que possible, une première version utilisateurs de la chaîne complète de traitement sera proposée sur le site www.univ-nancy2.fr/pers/namer.

Buvet, P-A, F. Moreau et M. Silberztein 2002 «INTEX et la recherche d'informations», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Corbin, D. 1987 *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, Niemeyer; 2e éd., 1991, Villeneuve d'Ascq, PUL.

Corbin, D. 1991 «Introduction : la formation des mots, structures et interprétations», *Lexique* 10, Villeneuve d'Ascq, PUL, p. 7-30.

Corbin, D. 1997 «Décrire un affixe dans un dictionnaire», dans G. Kleiber, M. Riegel et coll., *Les formes du sens. Études de linguistique française, médiévale et générale offertes à Robert Martin à l'occasion de ses 60 ans*, Louvain-la-Neuve / Paris, Duculot, p. 79-94.

Corbin, D. 2000a «Pour en finir avec la parasynthèse», dans G. Kleiber, J.-C. Pellat, C. Buridant et coll., *Mélanges de grammaire et de linguistique française en hommage au professeur Martin Riegel*.

Corbin, D. 2000b «French (Indo-European : Romance)», dans G. Booij, C. Lehmann, J. Mugdan et coll., *Morphology. A Handbook on Inflection and Word Formation*, Berlin / New-York, Walter de Gruyter.

Corbin, D. 2000c «Préfixes et suffixes : du sens aux catégories», *Journal of French Linguistic Studies* 11-1 : 41-69.

Daille, B., C. Fabre et P. Sébillot 2002 «Applications of Computational Morphology», dans *Many Morphologies,* Somerville (Mass.), Cascadilla Press, p. 210-234.

Dal, G. 1997a *Grammaire du suffixe -et(te),* Paris, Didier, coll. Érudition.

Dal, G. 1997b «Du principe d'unicité catégorielle au principe d'unicité sémantique : incidence sur la formalisation du lexique construit morphologiquement», dans P.-A. Buvet, S. Cardey, P. Greenfield, H. Madec et coll., *Actes du colloque international Fractal 1997, BULAG* numéro spécial, p. 105-115.

Darmoni, S. et coll. 2003 «VUMeF : Extending the French Involvement in the UMLS Metathesaurus», *AMIA 2003*, Washington, p.824.

Eggert, E., D. Maurel et O. Piton, 2002 «La formation des gentilés sur Internet», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Émirkanian, L. et E. Chieze 2002 «Variations morphologiques, sémantiques et RI sur le Web», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Fairon, C. 2000a «Parsing a Web Site as a Corpus», dans C. Fairon et coll., *Analyse lexicale et syntaxique : le système INTEX*, Amsterdam, Benjamins.

Fairon, C. 2000b «GlossaNet, un agent de veille. Utilisation de ressources linguistiques pour la recherche d'information sur le Web», dans Ch. Jacquemin et coll. *Traitement automatique des langues pour la recherche d'Information*, *Revue TAL* 41-2, Paris, Klincksieck.

Fouqueré, Ch. et F. Issac 2002 «Pertinence thématique de variations de requêtes», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Fourour N. et E. Morin 2002 «Apport du Web dans la reconnaissance d'entités nommées», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Fradin, B. 1994 «L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie», *Revue TAL* 35-2 : 9-48, Paris, Klincksieck.

Fradin, B. 2003 *Nouvelles approches en morphologie*. Paris, PUF.

Grabar,N. et S. Berland 2001 «Construire un corpus web pour l'acquisition terminologique», communication aux Journées TIA, Nancy.

Grabar, N. et P. Zweigenbaum, 2003 «Productivité à travers domaines et genres : dérivés adjectivaux et langue médicale», dans G. Dal et coll. «La productivité en questions et en expérimentations», *Langue Française* 140 : 102-125.

Grefenstette, G. et J. Nioche 2000 «Estimation of English and non-English language use on the WWW», dans *Proceedings of RIAO 2000 : Content-Based Multimedia Information Access*, p. 237-246, Paris, C.I.D.

Gruaz C. Ch. Jacquemin et E. Tzoukermann 1996, «Une approche à deux niveaux de la morphologie dérivationnelle du français», dans *Séminaire Lexique du GDR-PRC Communication Homme-Machine,* p. 107-114, Grenoble.

Hathout, N., F. Namer et G. Dal 2002 «An experimental Constructional Database : The MorTAL Project», dans P. Boucher et coll. *Many Morphologies*, Somerville (Mass.), Cascadilla Press, p. 178-209.

Hathout, N. et L. Tanguy 2002 «Vers une autodétection des webnéologismes», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse. [texte ici même]

Heiden, S. et P. Lafon 2002 «Lectures assistées de l'Encyclopédie électronique : philologic et Weblex», dans *Recherches sur Diderot et sur l'Encyclopédie*, p. 91-102.

Karttunen, L. 1983. «KIMMO : A general morphological processor.» *Linguistic Forum* 22 : 163-186.

Kerleroux, F. 1996 *La coupure invisible*, Presses Universitaires du Septentrion, Lille.

Kerleroux, F. 1997 «De la limitation de l'homonymie entre noms déverbaux convertis et apocopes de noms déverbaux suffixés», *Lexicales* 1 : 163-172, Lille.

Kerleroux, F. 2000 «Identification d'un procédé morphologique : la conversion», *Faits de Langue* 14 : 89-100.

Krott, A., H. Baayen et R. Schreuder 1999 «Complex words in complex words», *Linguistics* 37-5 : 905-926.

Lebarbé, T. 2002 «Validation des relations de dépendance par la cooccurrence sur Internet : présentation critique», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse.

Namer, F. 2000 «Flemm, un analyseur flexionnel du français à base de règles», *TAL* 41 –2 : 523-548, Paris, Hermès.

Namer, F. 2002 «Acquisition de sens à partir d'opérations morphologiques en français : étude de cas», *Actes de TALN 02,* Nancy, p.235-244.

Namer, F 2003a «WaliM : valider les unités morphologiques par le Web», dans *Silexicales «Les Unités morphologiques»,* Lille, p. 142-150.

Namer F. 2003b «Productivité morphologique, représentativité et complexité de la base : le système MoQuête», dans G. Dal et coll. *La productivité en questions et en expérimentations*, *Langue Française* 140 : 79-101.

Namer F. à paraître 2004a «Automatiser l'analyse morphosémantique non affixale : le système DériF», *Cahiers de Grammaire* 28, Toulouse.

Namer F. à paraître 2004b «Acquisizione automatica di semantica lessicale in francese: il sistema di trattamento computazionale della formazione delle parole DériF», dans *Atti del 27º Congresso Internazionale della Società di Linguistica Italiana*, L'Aquila.

Namer F. et P. Zweigenbaum à paraître 2004 «Acquiring meaning for French medical terminology: contribution of morphosemantics», dans *Proceedings of MEDINFO 2004*, San Francisco.

Plénat, M. et M. Roché 2000 «Prosodic constraints on suffixation in French». À paraître dans *Proceedings of the Third Mediterranean Morphology Meeting*, Barcelone.

Rey, A. et coll. 1998 *Le Robert- Dictionnaire Historique de la langue française,* Paris.

Schmid, H. 1994 «Probabilistic Part-of-Speech Tagging Using Decision Trees», dans *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, Manchester.

Sproat, R., 1992 *Morphology and Computation*, Cambridge (Mass.), MIT Press.

Tanguy, L. et Hathout, N. 2002 «Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web», *Actes de TALN02,* Nancy, p. 245-254.

Tazine, C. 2002 «Création automatique de modèle de langage n-grammes depuis Internet par une mesure de distance», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse.

Temple, M. 1996 *Pour une sémantique des mots construits*, Villeneuve d'Ascq, Presses Universitaires du Septentrion.

Torzec, N. 2002 «Construction d'un corpus électronique annoté dédié au traitement linguistique des messages électroniques», communication présentée au colloque TALN, Corpus et Web 2002, Villetaneuse.

Zweigenbaum, P. et coll. 2003a «Towards a Unified Medical Lexicon for French», dans *Proceedings of MIE 2003*, Saint-Malo, p. 415-420.

Zweigenbaum, P. et coll. 2003b «UMLF : a Unified Medical Lexicon for French», dans *Proceedings of AMIA 2003*, Washington, p. 1062.